Statistics for high-throughput experiments

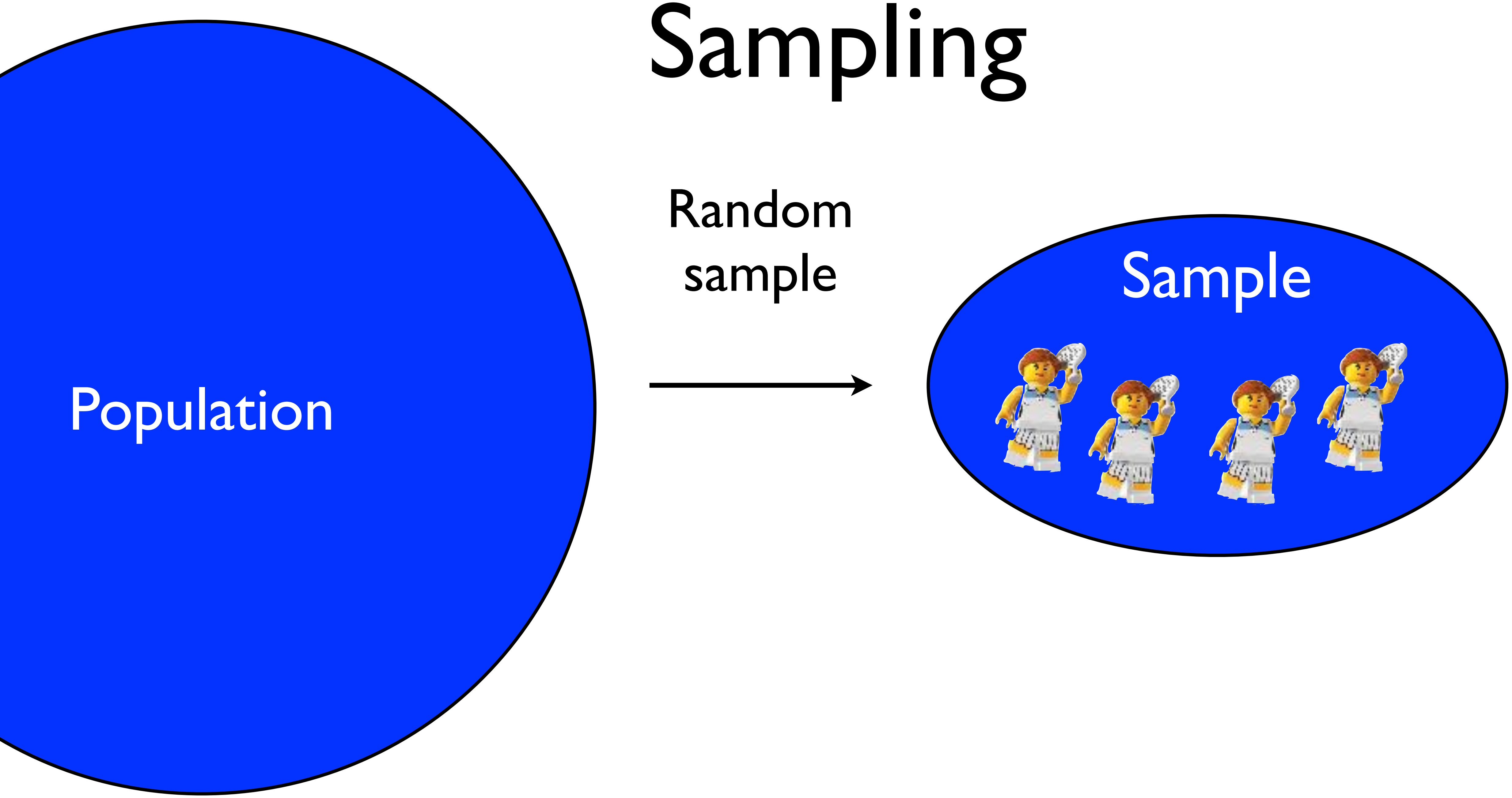
SciLifeLab

SERC
Swedish e-Science Research Centre

KTH
VETENSKAP
OCH KONST

Lukas Käll
KTH, School of Biotechnology
lukas.kall@scilifelab.se

Sampling



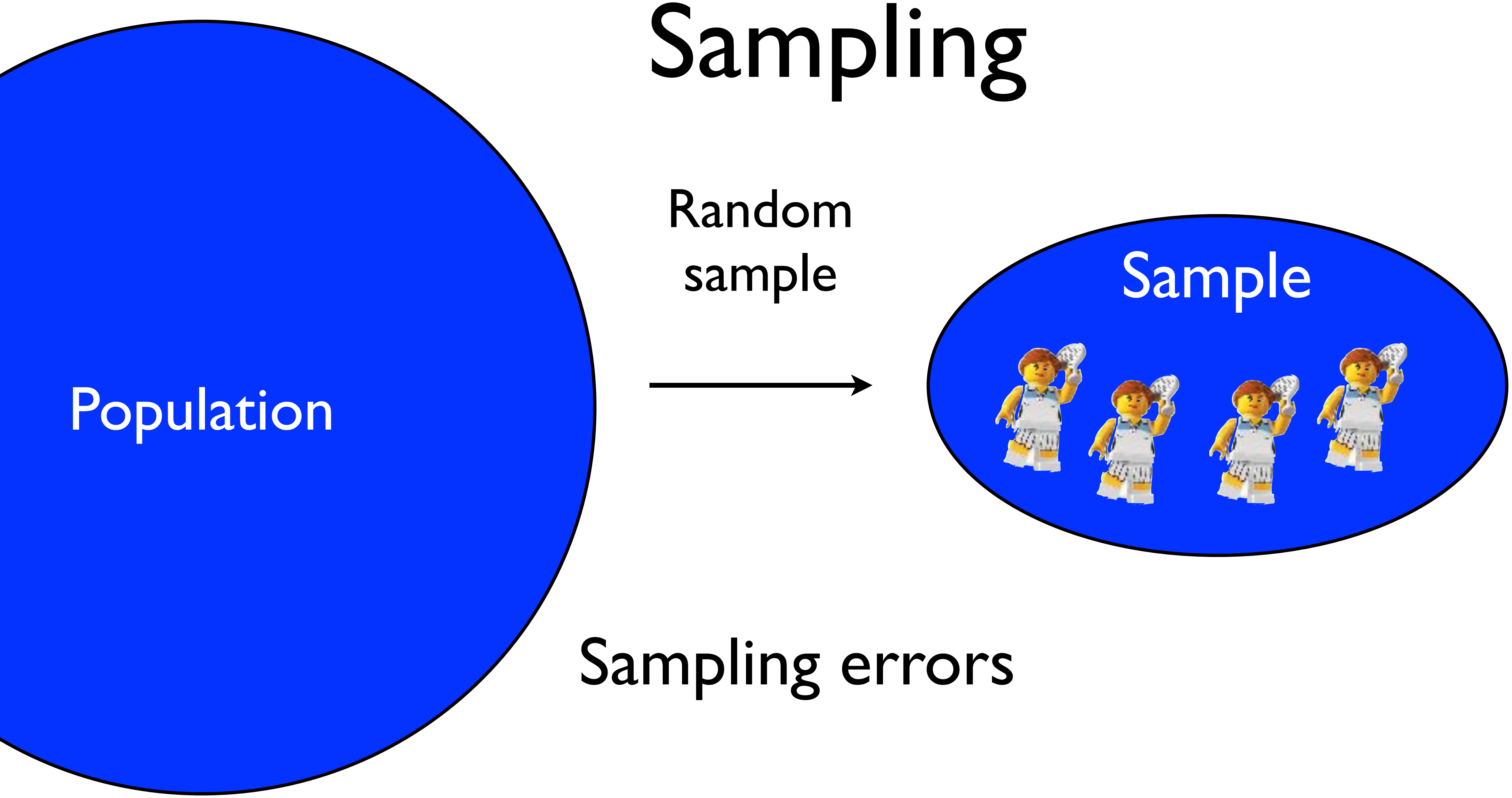
Population

Random sample

Sample

- We observe a population through a selected sample

Sampling



Population

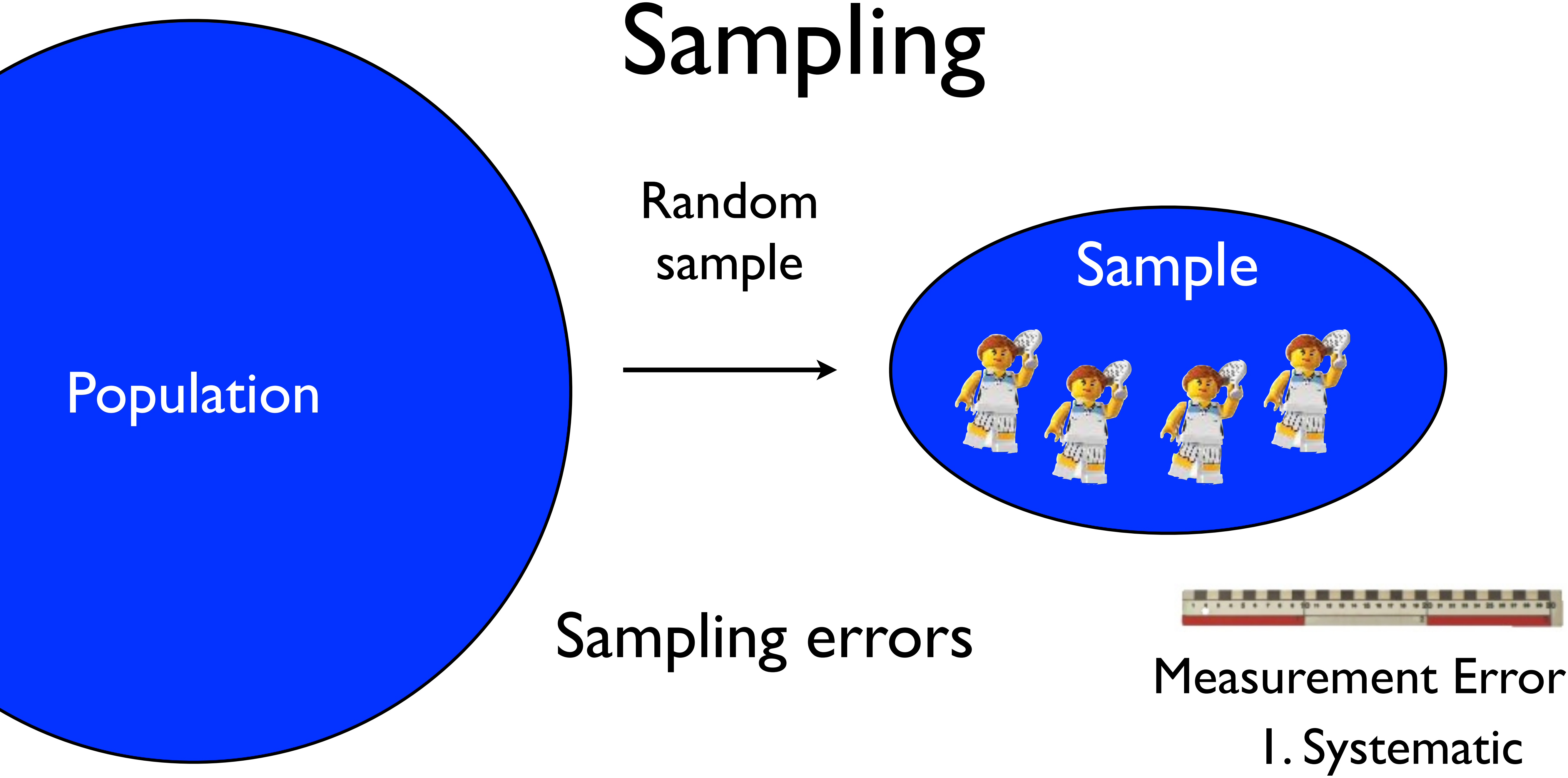
Random sample

Sample

Sampling errors

- We observe a population through a selected sample

Sampling



Population

Random sample

Sample

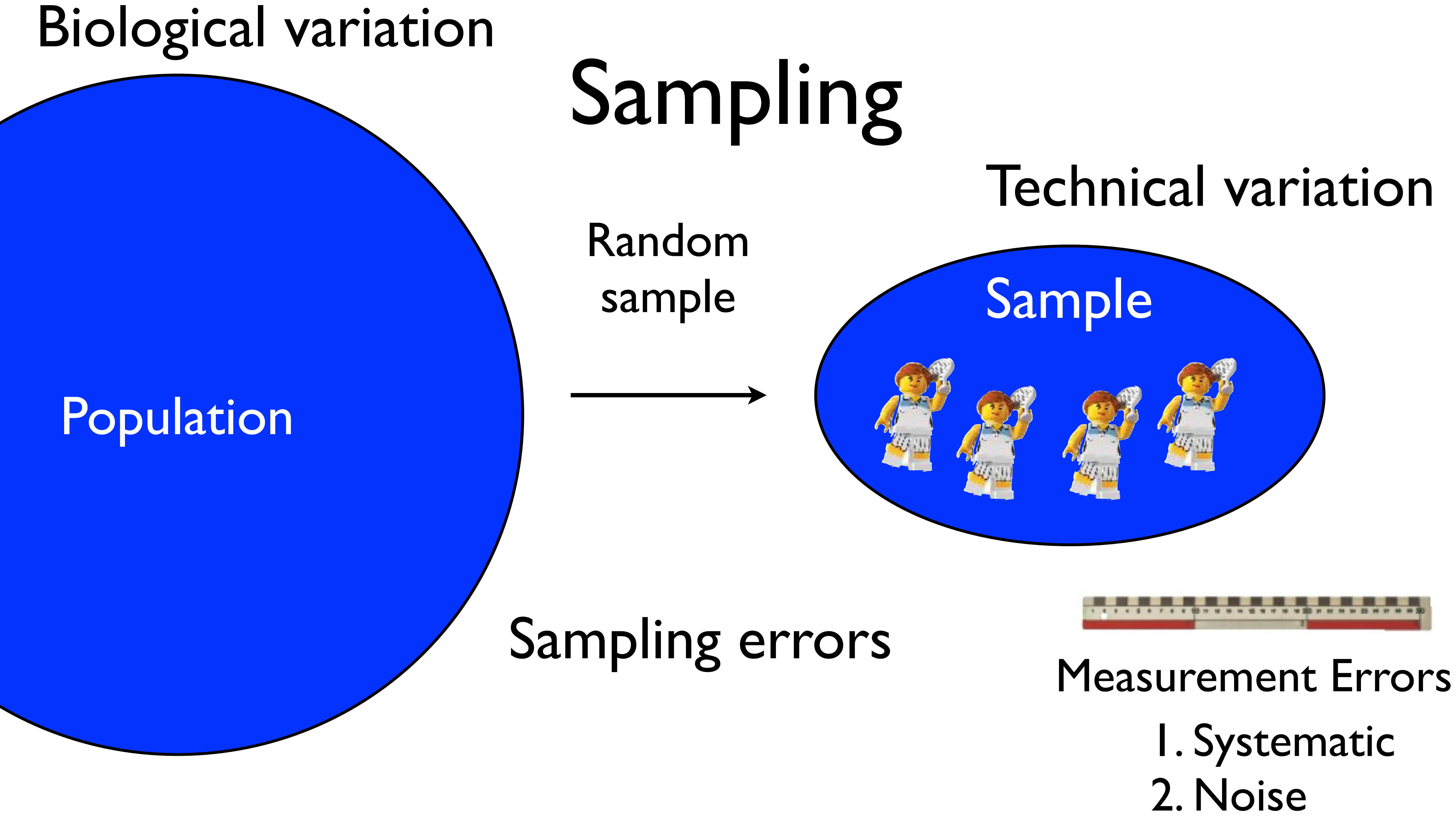
Sampling errors

Measurement Errors

1. Systematic
2. Noise

- We observe a population through a selected sample

Sampling



Biological variation

Population

Random sample

Sample

Technical variation

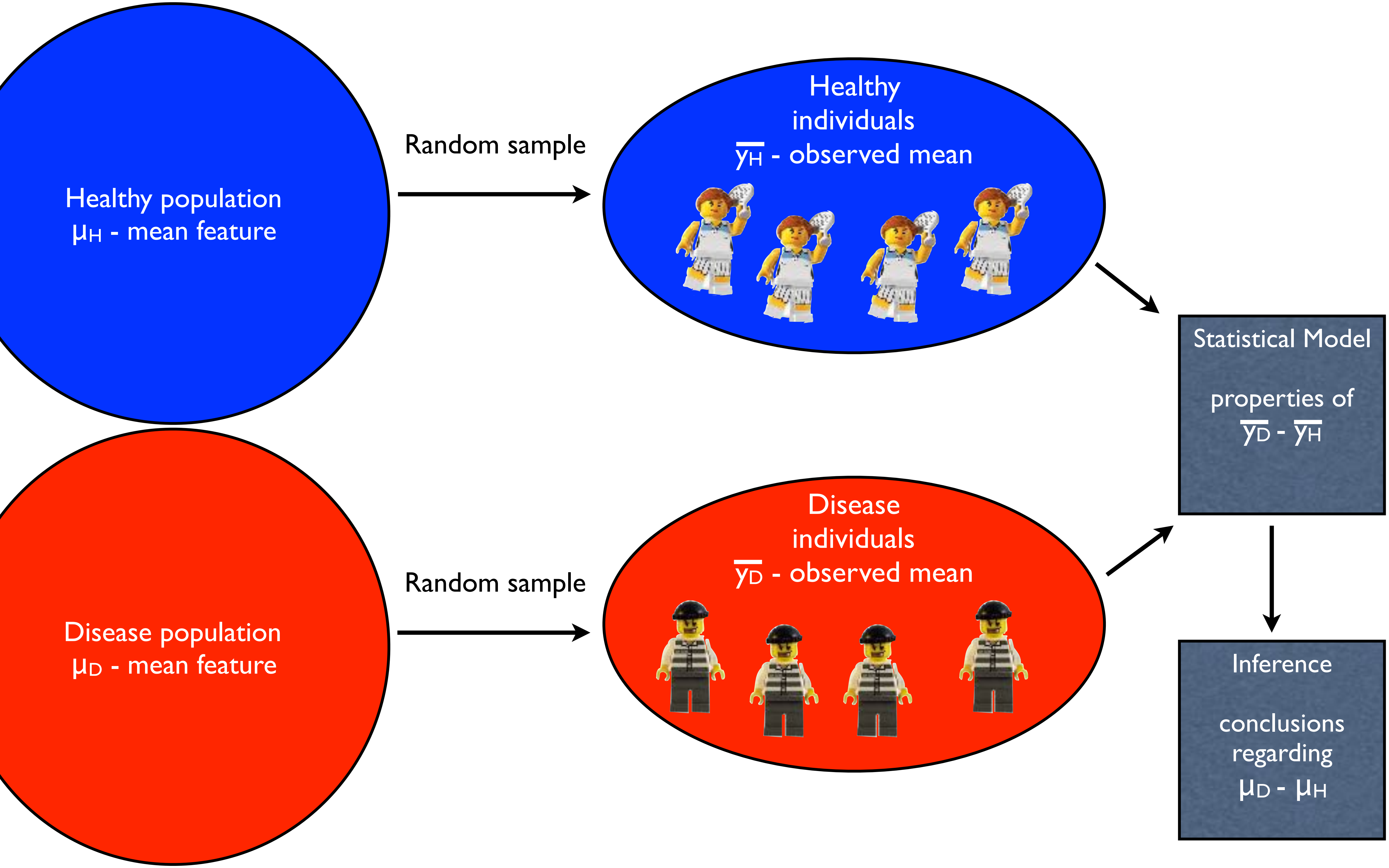
Sampling errors

Measurement Errors

1. Systematic
2. Noise

- We observe a population through a selected sample

Statistical inference procedure



Healthy population
μ_H - mean feature

Random sample

Healthy
individuals
$\overline{y}_H$ - observed mean

Disease population
μ_D - mean feature

Random sample

Disease
individuals
$\overline{y}_D$ - observed mean

Statistical Model

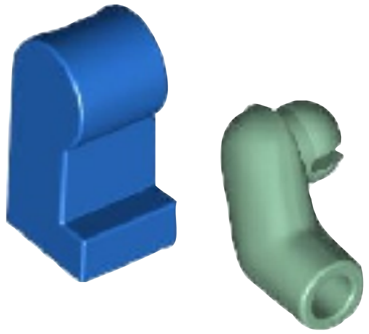
properties of
$\overline{y}_D - \overline{y}_H$

Inference

conclusions
regarding
$\mu_D - \mu_H$

Hypothesis testing

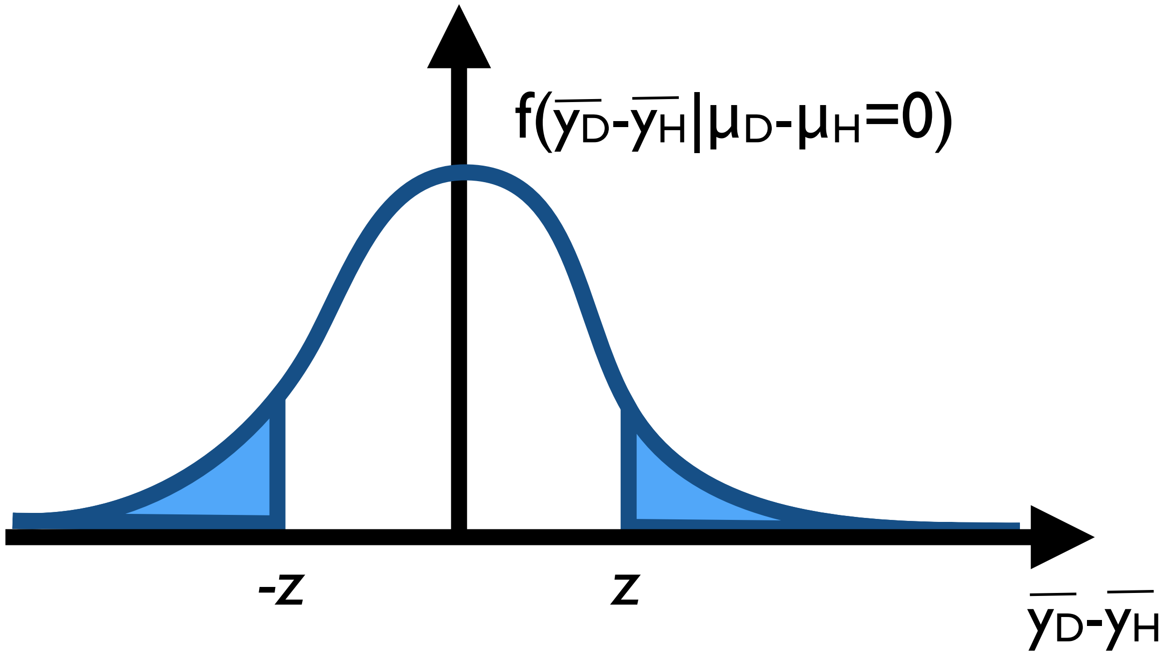
- *H_0*: The *null* hypothesis. The situation we are not interested in (typically μ_D-μ_H=0)



- *H_1*: The *alternative* hypothesis. The situation we want to detect (typically μ_D-$\mu_H \neq 0$)

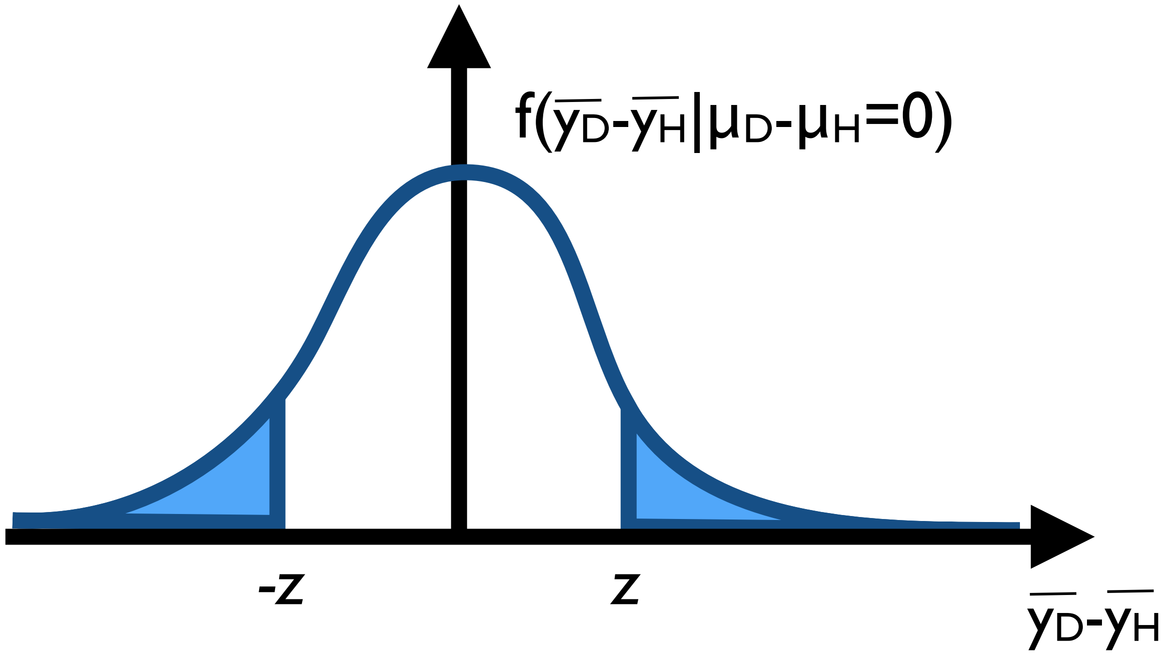
p value

- $\Pr(|\bar{y}_D - \bar{y}_H| \geq z \mid \mu_D - \mu_H = 0)$, *i.e.* the probability to a result at least as extreme as the one that was observed given H_0.

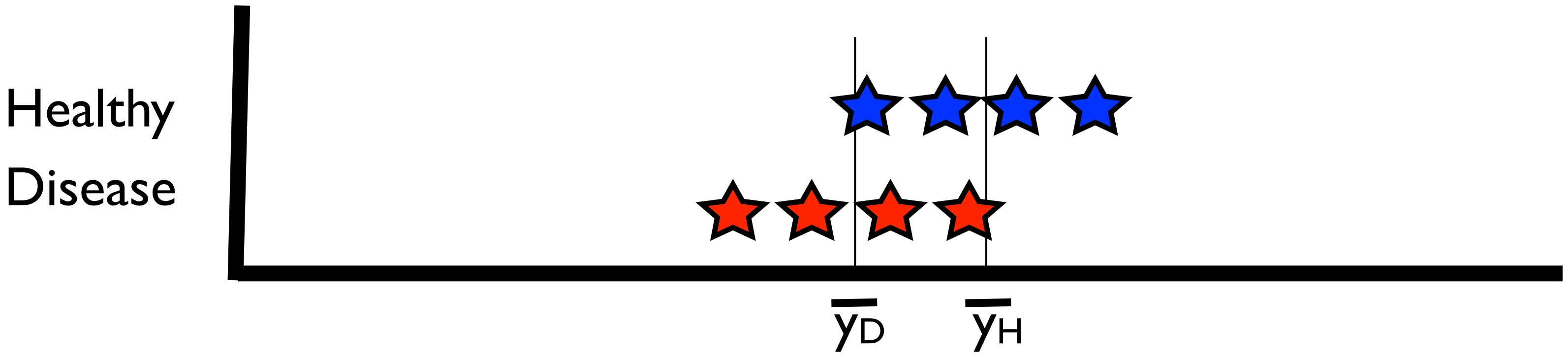


p value

- $\Pr(|\bar{y}_D - \bar{y}_H| \geq z | \mu_D - \mu_H = 0)$, *i.e.* the probability to a result at least as extreme as the one that was observed given H_0.
- *p* values are uniformly distributed under H_0.



Student’s t-test



Difference between sample means

$$t= \frac{\bar{y}_D-\bar{y}_H}{SE(\bar{y}_D-\bar{y}_H)}$$

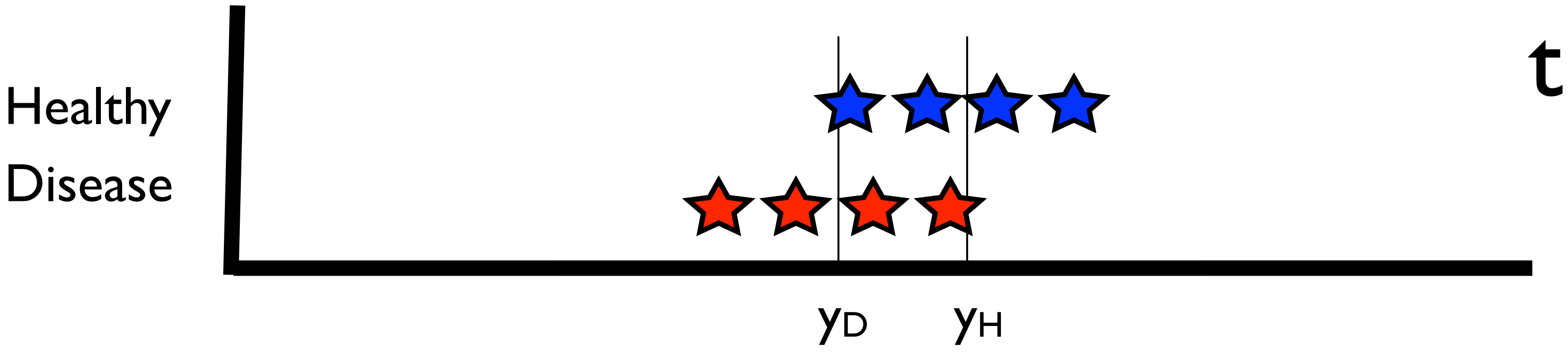
Standard error of difference of means

$$SE(\bar{y}_D-\bar{y}_H) \propto n^{-1/2}$$

Larger t ⇨ More significance

⇨ Lower p-value

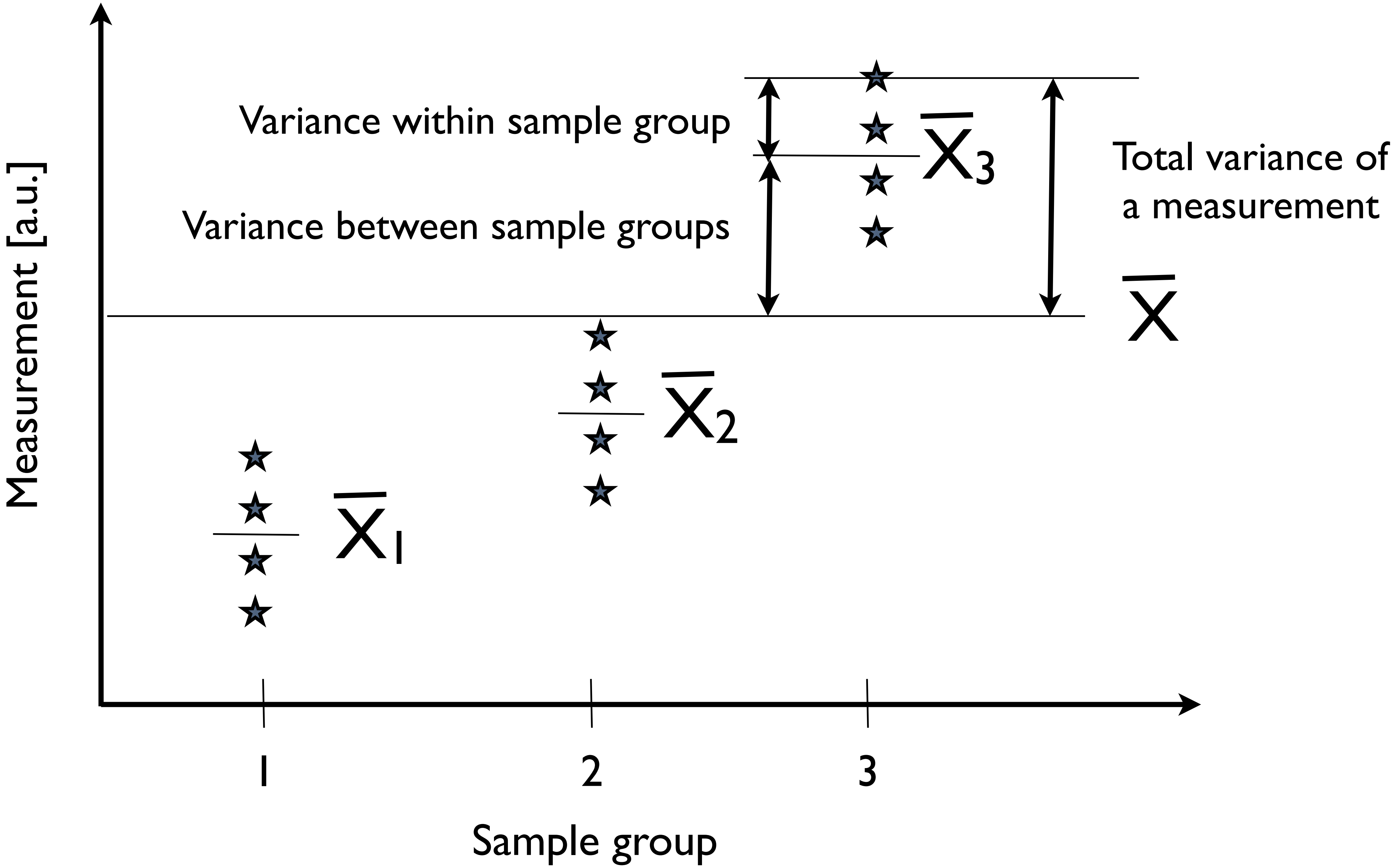
Student's t-test



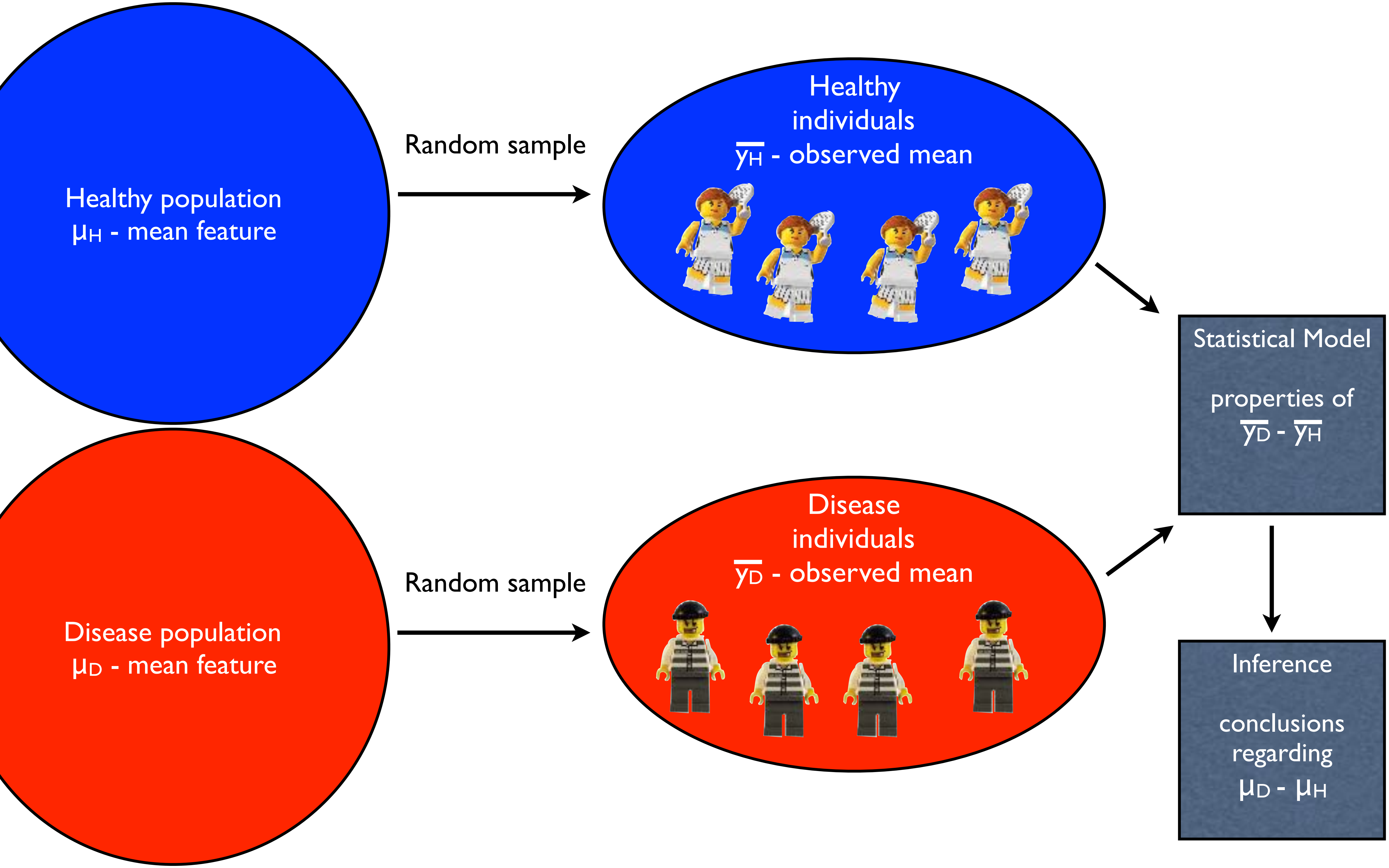
$$t = \frac{\bar{y}_D - \bar{y}_H}{SE(\bar{y}_D - \bar{y}_H)}$$

- Assumes that
  - the populations features and the errors follow normal distributions
- Variants include possibilities to test under
  - unequal sample sizes, unequal population variance, paired samples

ANalysis Of VAriance (ANOVA)



Statistical inference procedure



Healthy population
μ_H - mean feature

Random sample

Healthy
individuals
$\overline{y}_H$ - observed mean

Disease population
μ_D - mean feature

Random sample

Disease
individuals
$\overline{y}_D$ - observed mean

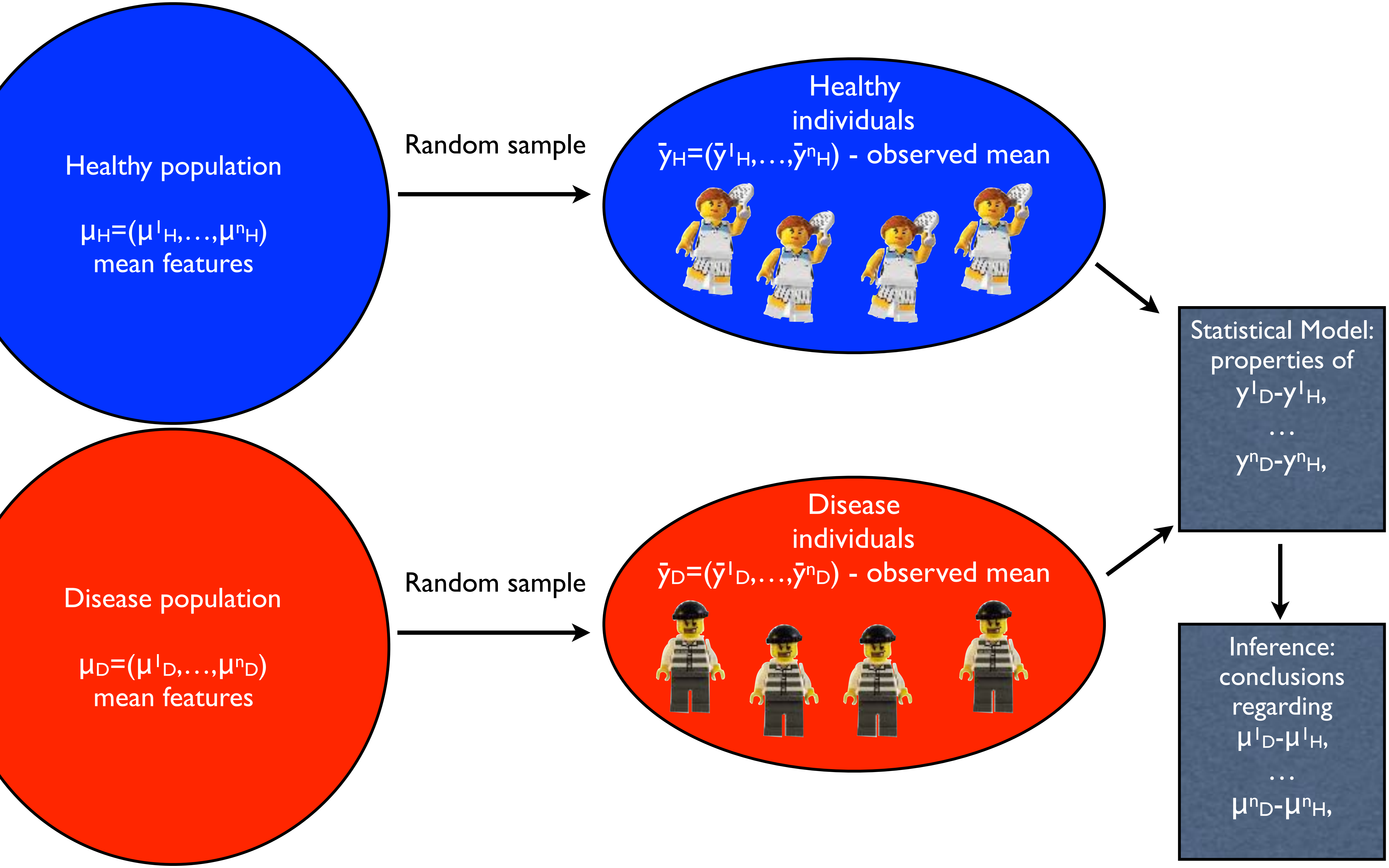
Statistical Model

properties of
$\overline{y}_D$ - $\overline{y}_H$

Inference

conclusions
regarding
μ_D - μ_H

Multiple measurements per sampled individual



Healthy population

$\mu_H=(\mu^1_H,\ldots,\mu^n_H)$
mean features

Random sample

Healthy individuals
$\bar{y}_H=(\bar{y}^1_H,\ldots,\bar{y}^n_H)$ - observed mean

Disease population

$\mu_D=(\mu^1_D,\ldots,\mu^n_D)$
mean features

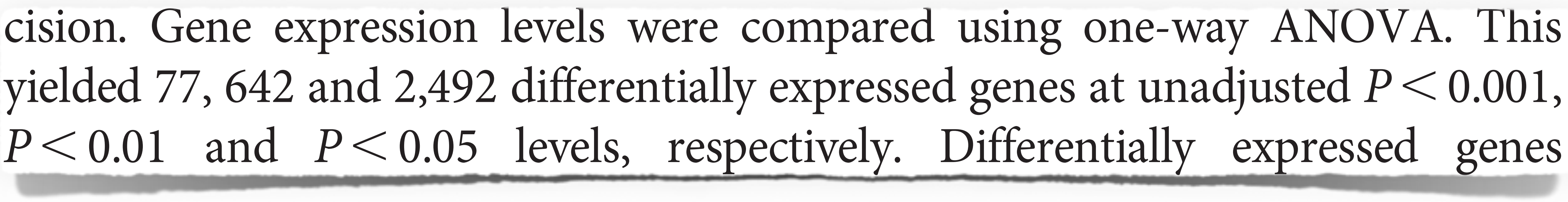
Random sample

Disease individuals
$\bar{y}_D=(\bar{y}^1_D,\ldots,\bar{y}^n_D)$ - observed mean

Statistical Model: properties of
$y^1_D-y^1_H$,
…
$y^n_D-y^n_H$,

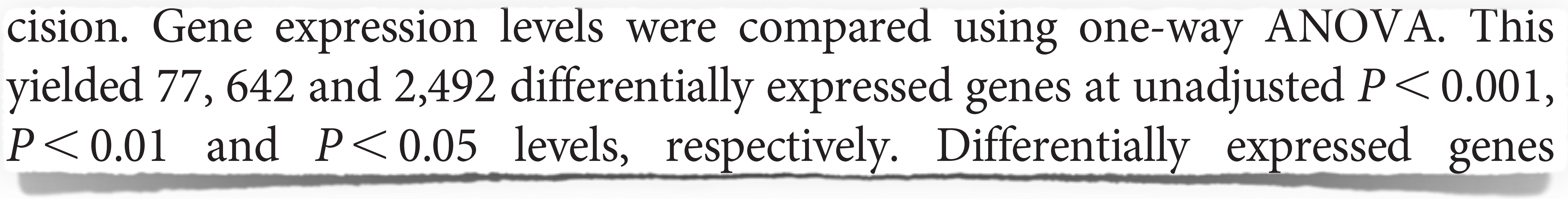
Inference: conclusions regarding
$\mu^1_D-\mu^1_H$,
…
$\mu^n_D-\mu^n_H$,

Motivating Example: micro Array study (published in Nature)



cision. Gene expression levels were compared using one-way ANOVA. This yielded 77, 642 and 2,492 differentially expressed genes at unadjusted $P < 0.001$, $P < 0.01$ and $P < 0.05$ levels, respectively. Differentially expressed genes

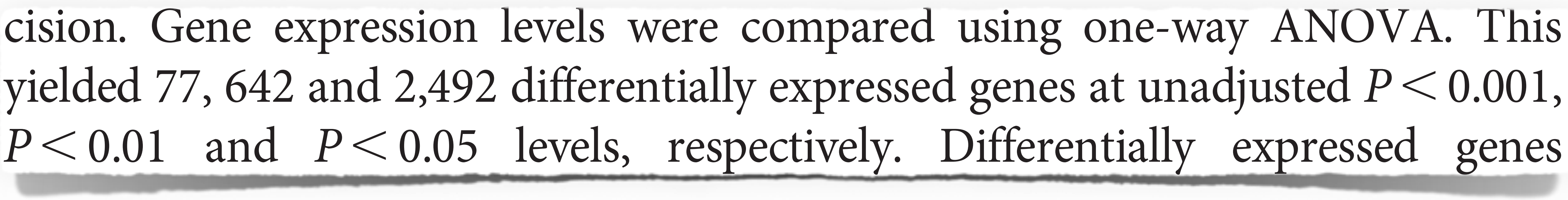
How many of 50 000 probes would we expect to be significant under the null hypothesis?



cision. Gene expression levels were compared using one-way ANOVA. This yielded 77, 642 and 2,492 differentially expressed genes at unadjusted $P < 0.001$, $P < 0.01$ and $P < 0.05$ levels, respectively. Differentially expressed genes

How many of 50 000 probes would we expect to be significant under the null hypothesis?

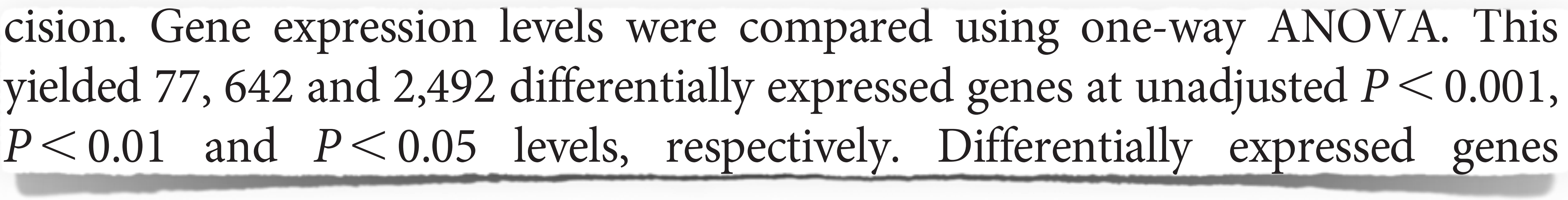
with P<0.001: 50000*0.001= 50



cision. Gene expression levels were compared using one-way ANOVA. This yielded 77, 642 and 2,492 differentially expressed genes at unadjusted $P < 0.001$, $P < 0.01$ and $P < 0.05$ levels, respectively. Differentially expressed genes

How many of 50 000 probes would we expect to be significant under the null hypothesis?

with P<0.001: 50000*0.001= 50
with P<0.01: 50000*0.01 = 500
with P<0.05: 50000*0.05 = 2500



cision. Gene expression levels were compared using one-way ANOVA. This yielded 77, 642 and 2,492 differentially expressed genes at unadjusted $P < 0.001$, $P < 0.01$ and $P < 0.05$ levels, respectively. Differentially expressed genes

Multiple Hypothesis Corrections

- Measures like p value accounts for the situation where we conduct one hypothesis test

Multiple Hypothesis Corrections

- Measures like p value accounts for the situation where we conduct one hypothesis test
- Simplest possible compensation: Bonferroni correction: divide your anticipated "familywise error rate" with the number of tests.
  e.g. for a "familywise error rate" threshold of 0.05 in an experiment with 50000 features we threshold individual *p* values with 0.05/50000=1E-6

  - Bonnferoni corrections are extremely conservative

Multiple Hypothesis Corrections

- Measures like p value accounts for the situation where we conduct one hypothesis test
- Simplest possible compensation: Bonferroni correction: divide your anticipated "familywise error rate" with the number of tests.
  e.g. for a "familywise error rate" threshold of 0.05 in an experiment with 50000 features we threshold individual *p* values with 0.05/50000=1E-6

  - Bonnferoni corrections are extremely conservative
- Better way: control for false discovery rate (FDR)

False Discovery Rate

score	type
0.0001	alternative (H_1)
0.00015	alternative (H_1)
0.00017	alternative (H_1)
0.0002	alternative (H_1)
0.00022	null (H_0)
0.00023	alternative (H_1)
0.00034	alternative (H_1)
0.00042	alternative (H_1)
0.00046	null (H_0)
0.00055	alternative (H_1)
0.00065	null (H_0)
0.00073	alternative (H_1)
0.00084	null (H_0)
...	...

threshold

False Discovery Rate

score	type
0.0001	alternative (H_1)
0.00015	alternative (H_1)
0.00017	alternative (H_1)
0.0002	alternative (H_1)
0.00022	null (H_0)
0.00023	alternative (H_1)
0.00034	alternative (H_1)
0.00042	alternative (H_1)
0.00046	null (H_0)
0.00055	alternative (H_1)
0.00065	null (H_0)
0.00073	alternative (H_1)
0.00084	null (H_0)
...	...

threshold

$$\frac{2}{10}$$

FDR(x) is the expectation value of the fraction of tests below threshold *x* that are generated under the null hypothesis

Mixture model

- We are studying a number of differences in feature means, some generated under the alternative hypothesis (H_1) and some to generated under the null hypothesis (H_0).

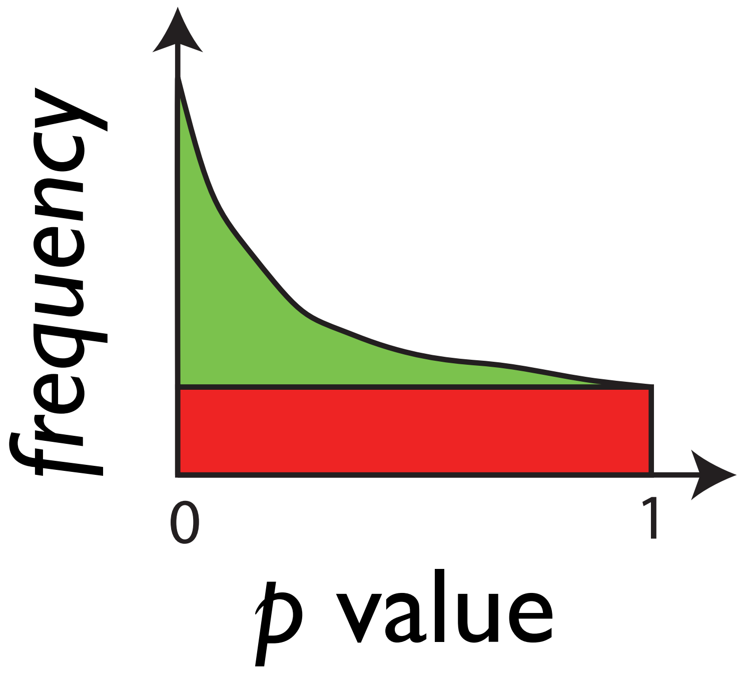
$$\Pr(p=t) = \Pr(\mathrm{H}=H_0)\Pr(p=t|\mathrm{H}=H_0) + \Pr(\mathrm{H}=H_1)\Pr(p=t|\mathrm{H}=H_1)$$

$$f(t)=\pi_0 f_0(t)+ \pi_1 f_1(t)$$

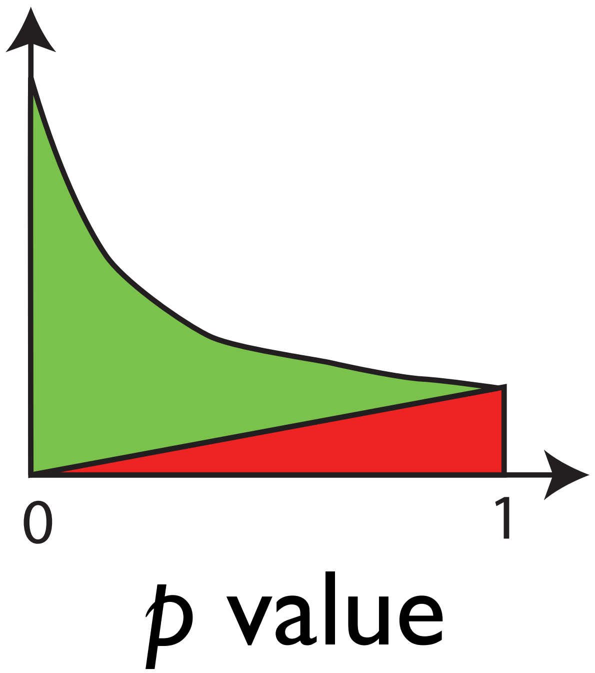
Concept test: distribution of *p* values

Which of the following histograms would be a likely outcome from a well calibrated high throughput experiment?

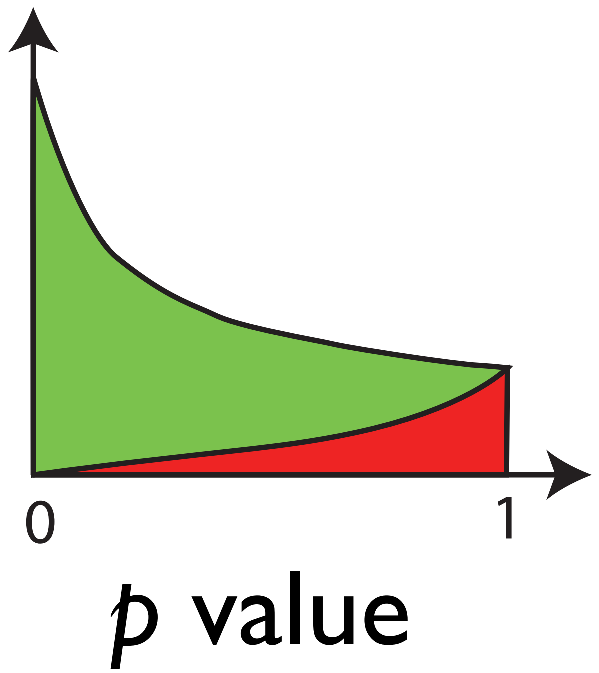
A



B



C

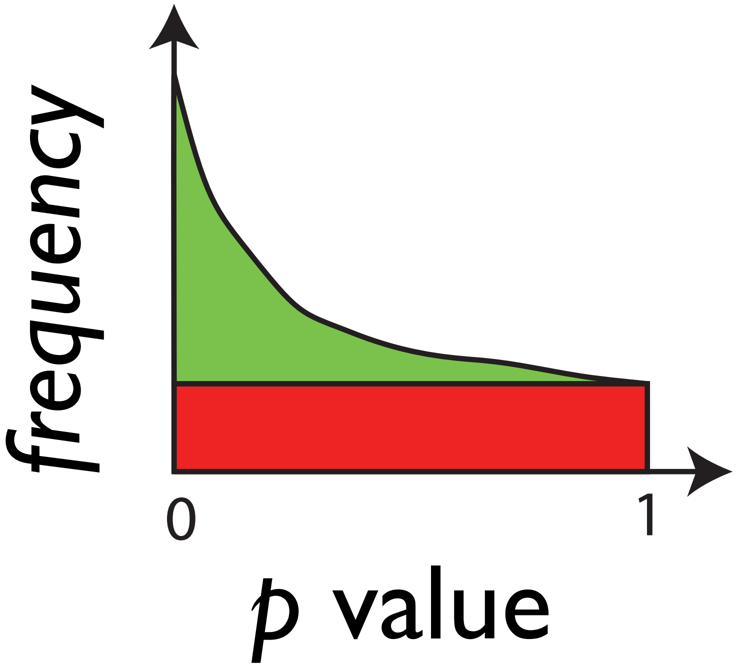


🟢 H_1 🔴 H_0

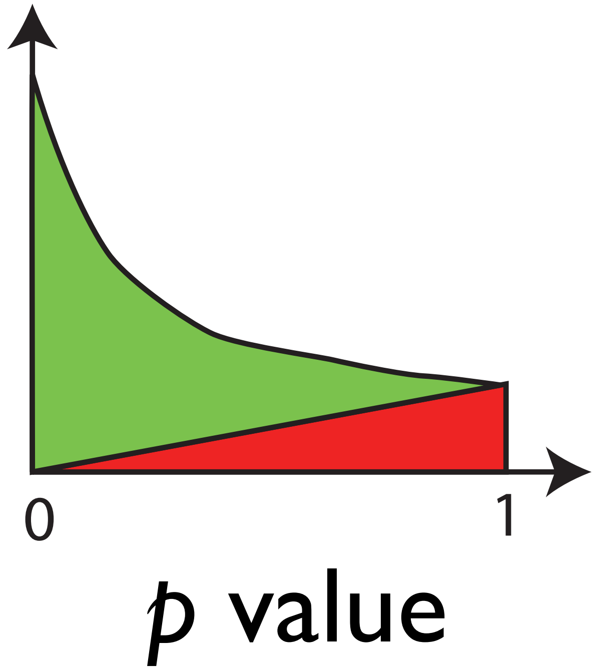
Concept test: distribution of *p* values

Which of the following histograms would be a likely outcome from a well calibrated high throughput experiment?

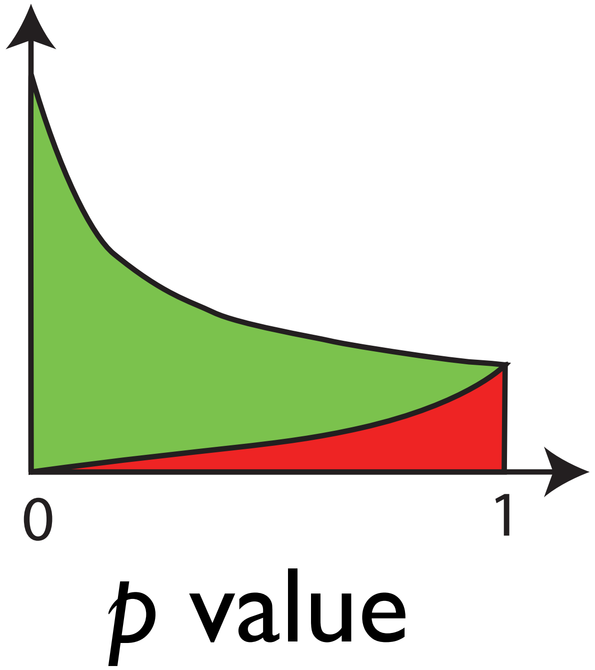
[✓ A]



B



C

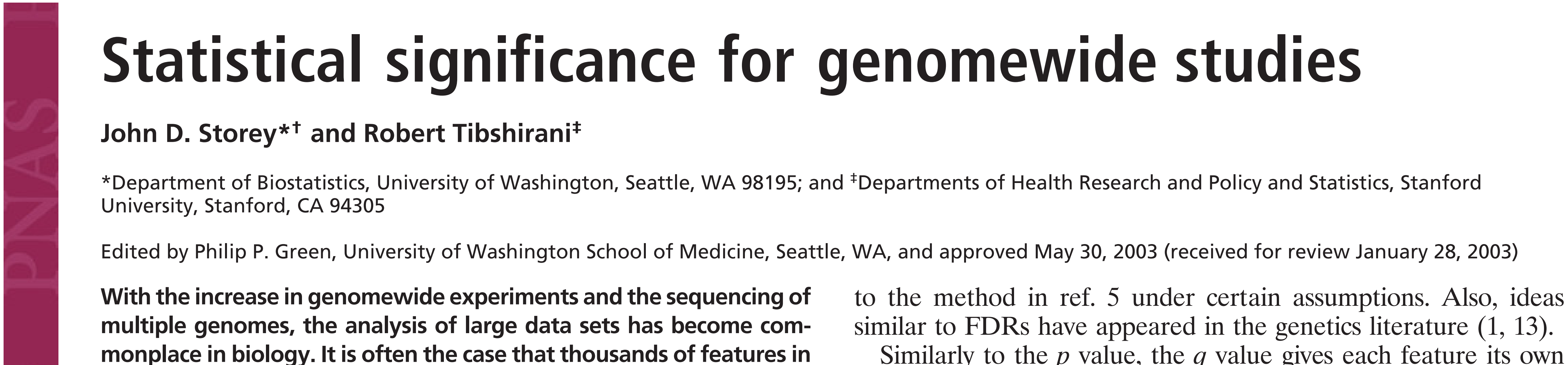


🟩 H_1 🟥 H_0

	Called significant	Called not significant	Total
Null true	F	$m_0 - F$	m_0
Alternative true	T	$m_1 - T$	m_1
Total	S	$m - S$	m

idéa [Benjamini and Hochberg 1995] - control for:

$$\frac{\text{no. false positive features}}{\text{no. significant features}} = \frac{F}{F+T} = \frac{F}{S}$$’



Statistical significance for genomewide studies

John D. Storey*[†] and Robert Tibshirani[‡]

*Department of Biostatistics, University of Washington, Seattle, WA 98195; and [‡]Departments of Health Research and Policy and Statistics, Stanford University, Stanford, CA 94305

Edited by Philip P. Green, University of Washington School of Medicine, Seattle, WA, and approved May 30, 2003 (received for review January 28, 2003)

With the increase in genomewide experiments and the sequencing of multiple genomes, the analysis of large data sets has become commonplace in biology. It is often the case that thousands of features in

to the method in ref. 5 under certain assumptions. Also, ideas similar to FDRs have appeared in the genetics literature (1, 13).

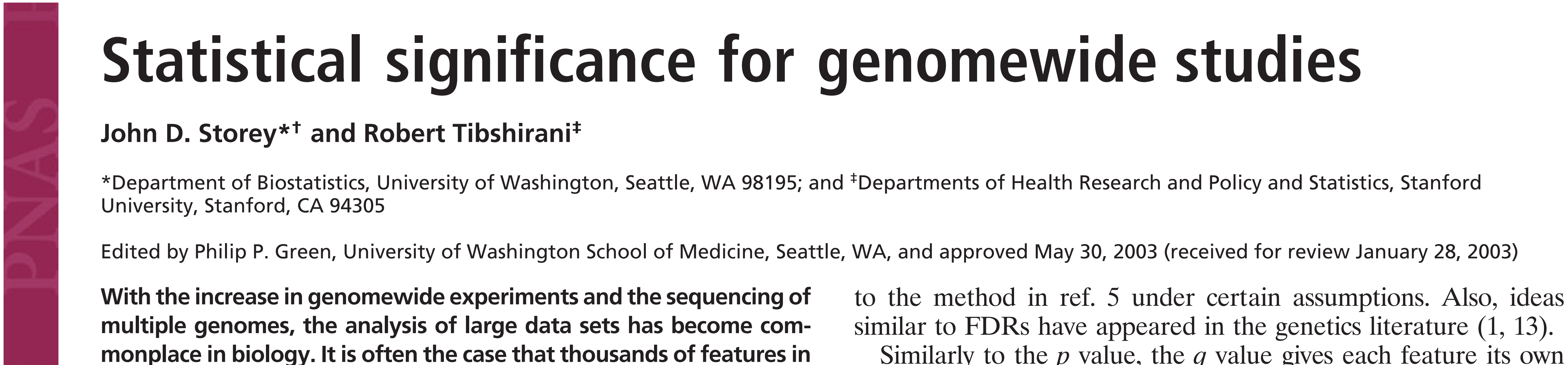
Similarly to the p value, the q value gives each feature its own

	Called significant	Called not significant	Total
Null true	F	$m_0 - F$	m_0
Alternative true	T	$m_1 - T$	m_1
Total	S	$m - S$	m

idéa [Benjamini and Hochberg 1995] - control for:

$$\frac{\text{no. false positive features}}{\text{no. significant features}} = \frac{F}{F+T} = \frac{F}{S}\text{'}$$

$$\text{FDR} = \text{E}\left[\frac{F}{F+T}\right] = \text{E}\left[\frac{F}{S}\right].$$



PNAS

Statistical significance for genomewide studies

John D. Storey*† and Robert Tibshirani‡

*Department of Biostatistics, University of Washington, Seattle, WA 98195; and ‡Departments of Health Research and Policy and Statistics, Stanford University, Stanford, CA 94305

Edited by Philip P. Green, University of Washington School of Medicine, Seattle, WA, and approved May 30, 2003 (received for review January 28, 2003)

With the increase in genomewide experiments and the sequencing of multiple genomes, the analysis of large data sets has become commonplace in biology. It is often the case that thousands of features in

to the method in ref. 5 under certain assumptions. Also, ideas similar to FDRs have appeared in the genetics literature (1, 13).

Similarly to the p value, the q value gives each feature its own

We got *m p* values, $p_1, p_2, \ldots, p_m$

for a threshold t we may say that:

$$F(t) = \#\,\{\text{null}\, p_i \leq t; i = 1, \ldots, m\} \text{ and}$$

$$S(t) = \#\,\{p_i \leq t; i = 1, \ldots, m\}.$$

$$\text{FDR}(t) = \text{E}\left[\frac{F(t)}{S(t)}\right].$$

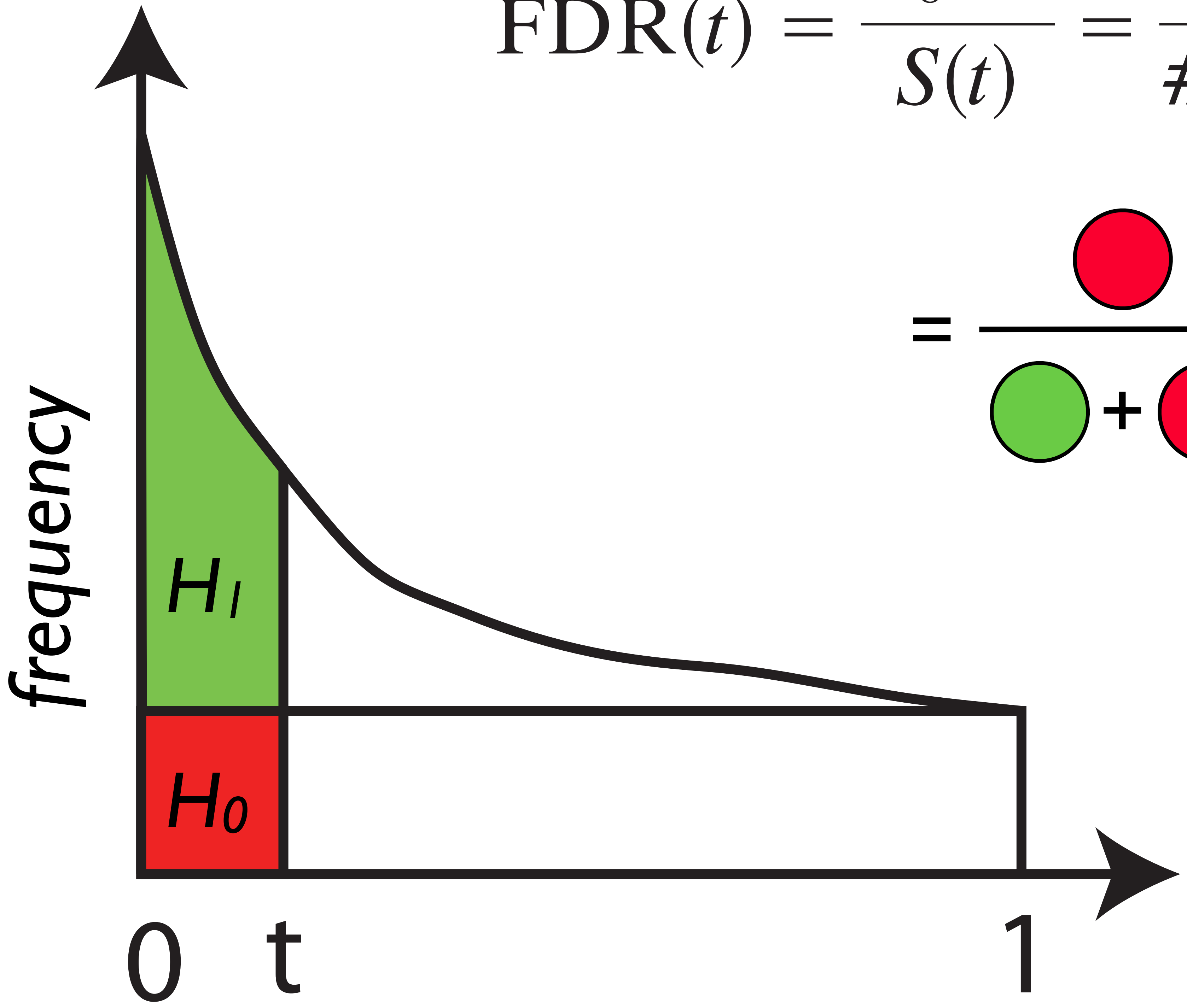
Evenly distributed *p* values: $F(t) = m_0 t = \pi_0 m t$

$$\widehat{\text{FDR}}(t) = \frac{\hat{\pi}_0 m \cdot t}{S(t)} = \frac{\hat{\pi}_0 m \cdot t}{\#\{p_i \leq t\}}.$$

Illustration of $\widehat{\text{FDR}}$

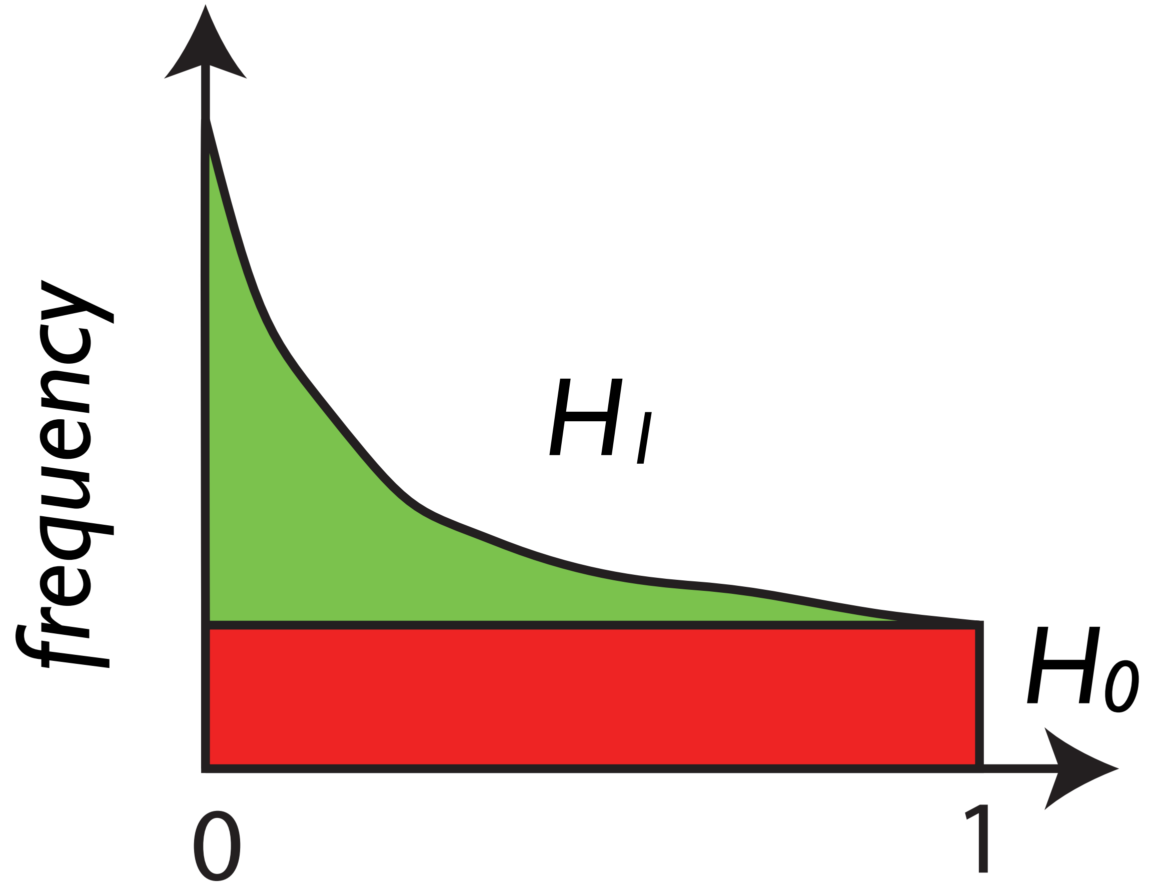
$$\widehat{\text{FDR}}(t) = \frac{\hat{\pi}_0 m \cdot t}{S(t)} = \frac{\hat{\pi}_0 m \cdot t}{\#\{p_i \leq t\}}.$$

= 🔴 / (🟢 + 🔴)



π_0

π_0 is the prior probability that a statistic is derived under H_0 *i.e.* Pr(H=H_0)

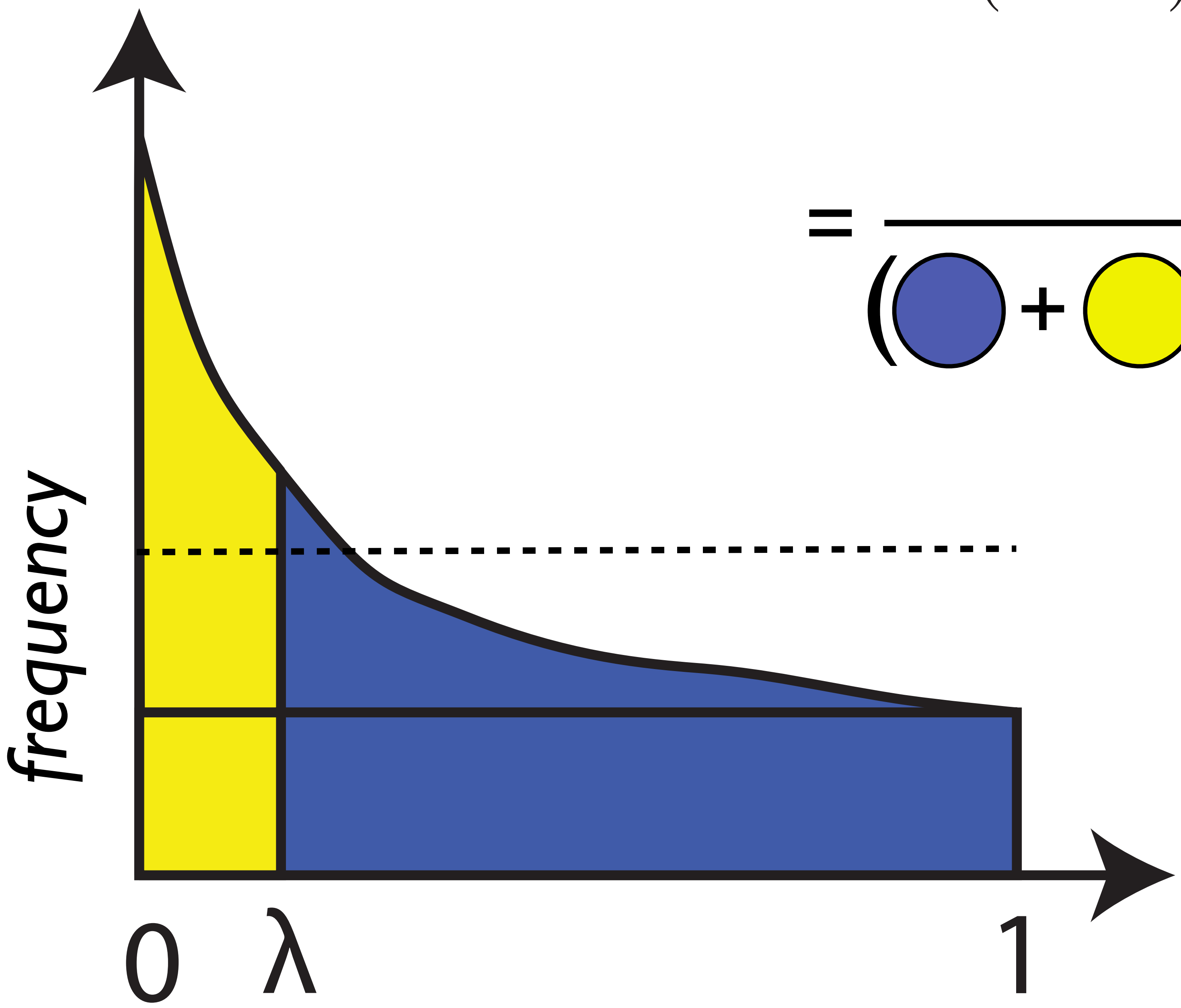


$$\pi_0 = \frac{🔴}{🟢 + 🔴}$$

π₀ estimation

$$\hat{\pi}_0(\lambda) = \frac{\# \{p_i > \lambda; i = 1, \ldots, m\}}{m(1-\lambda)},$$

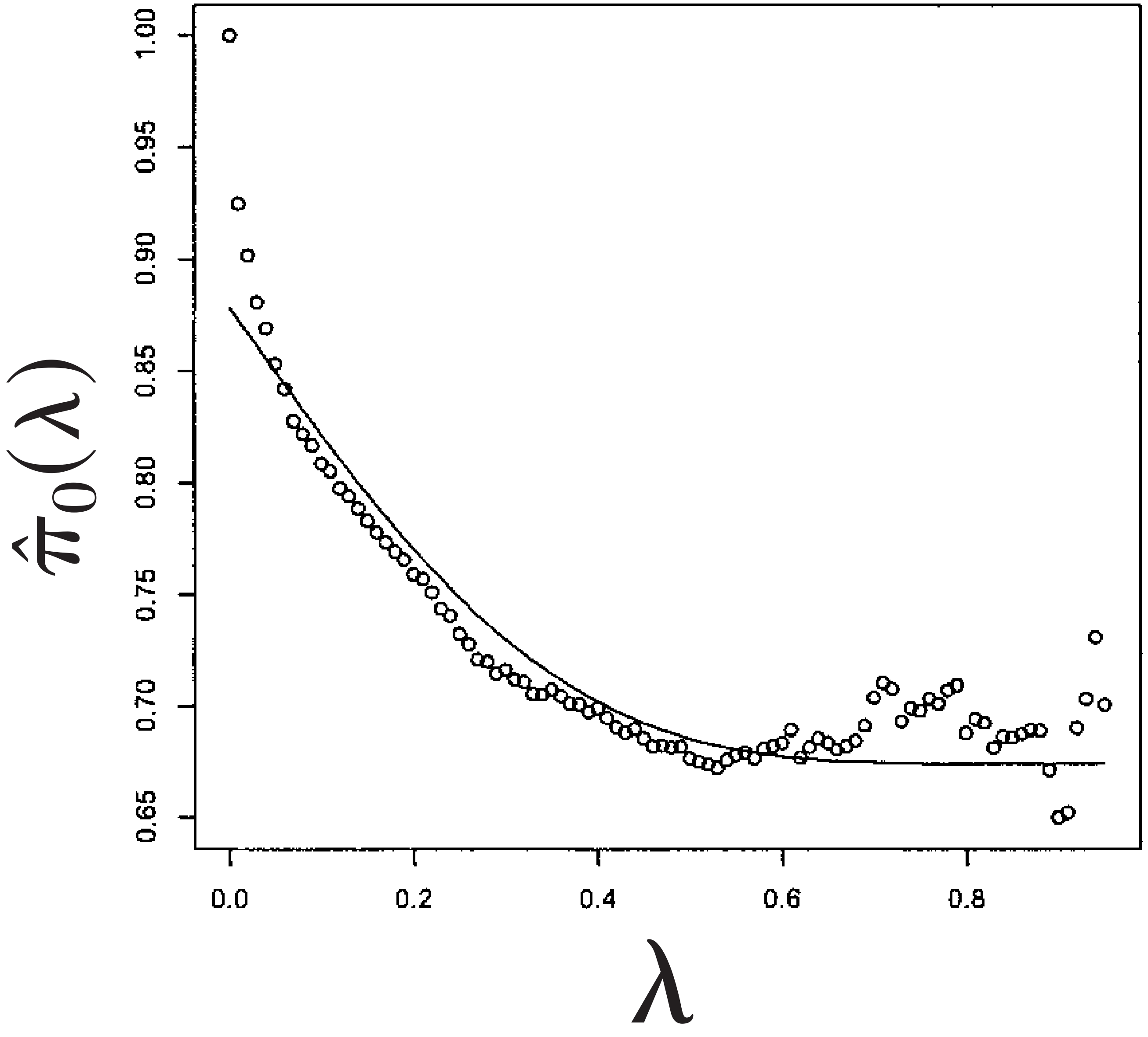
$$= \frac{🔵}{(🔵 + 🟡)(1-\lambda)}$$



π_0 estimation

Investigate the higher (close to 1) *p* values

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1, \ldots, m\}}{m(1-\lambda)},$$



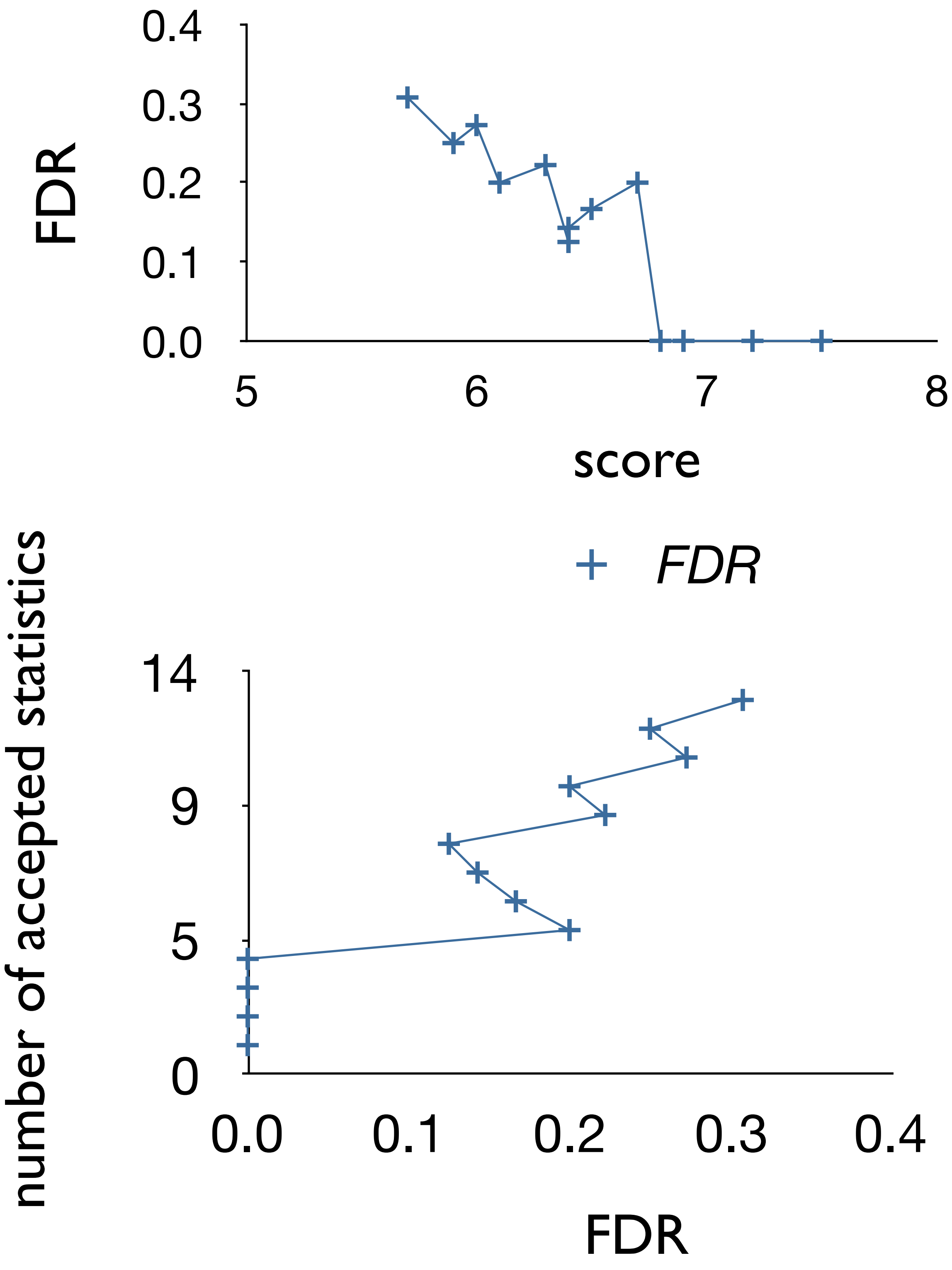
q value

A relevant measures to individual identifications that ensures monotonically increasing function with the *p* value threshold. The *q* value is defined as

$$\hat{q}(p_i) = \min_{t \geq p_i} \widehat{\mathrm{FDR}}(t).$$

q value

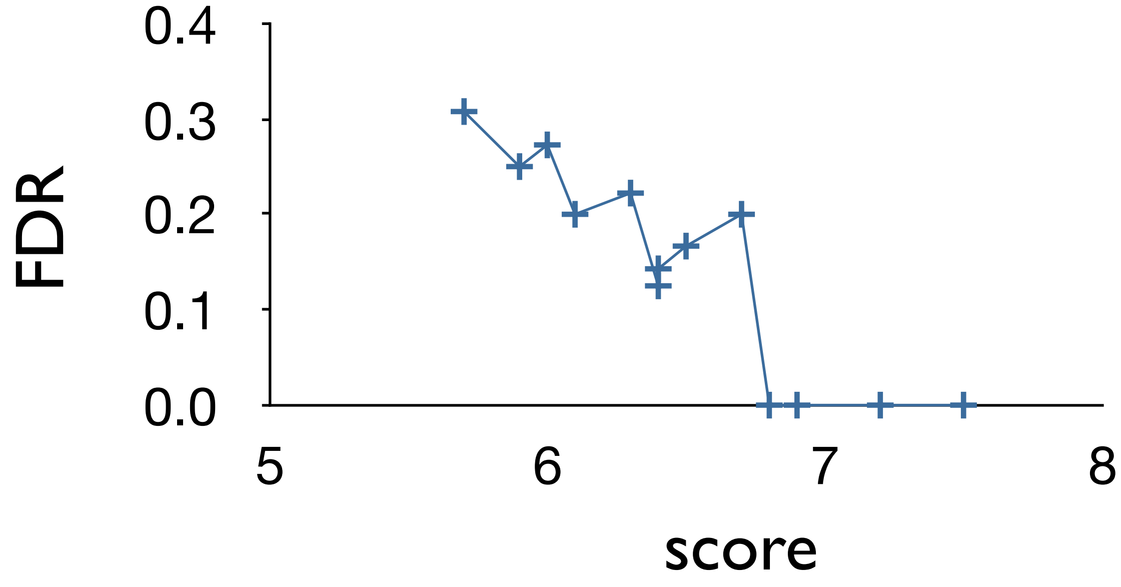
score	type
7.5	correct
7.2	correct
6.9	correct
6.8	correct
6.7	incorrect
6.5	correct
6.4	correct
6.4	correct
6.3	incorrect
6.1	correct
6	incorrect
5.9	correct
5.7	incorrect
...	...

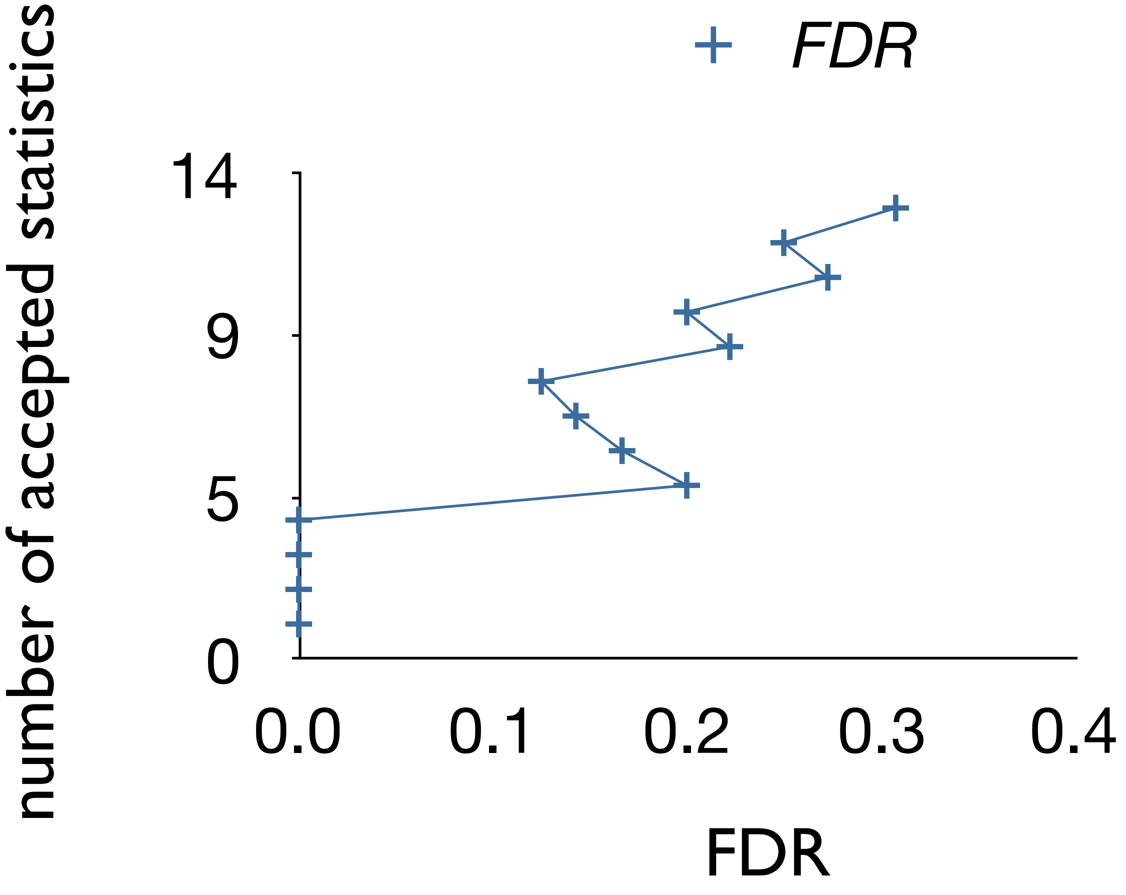


$$q(x)=\min_{x\geq x'}\{\mathrm{FDR}(x')\}$$

q value

score	type
7.5	correct
7.2	correct
6.9	correct
6.8	correct
6.7	incorrect
6.5	correct
6.4	correct
6.4	correct
6.3	incorrect
6.1	correct
6	incorrect
5.9	correct
5.7	incorrect
...	...

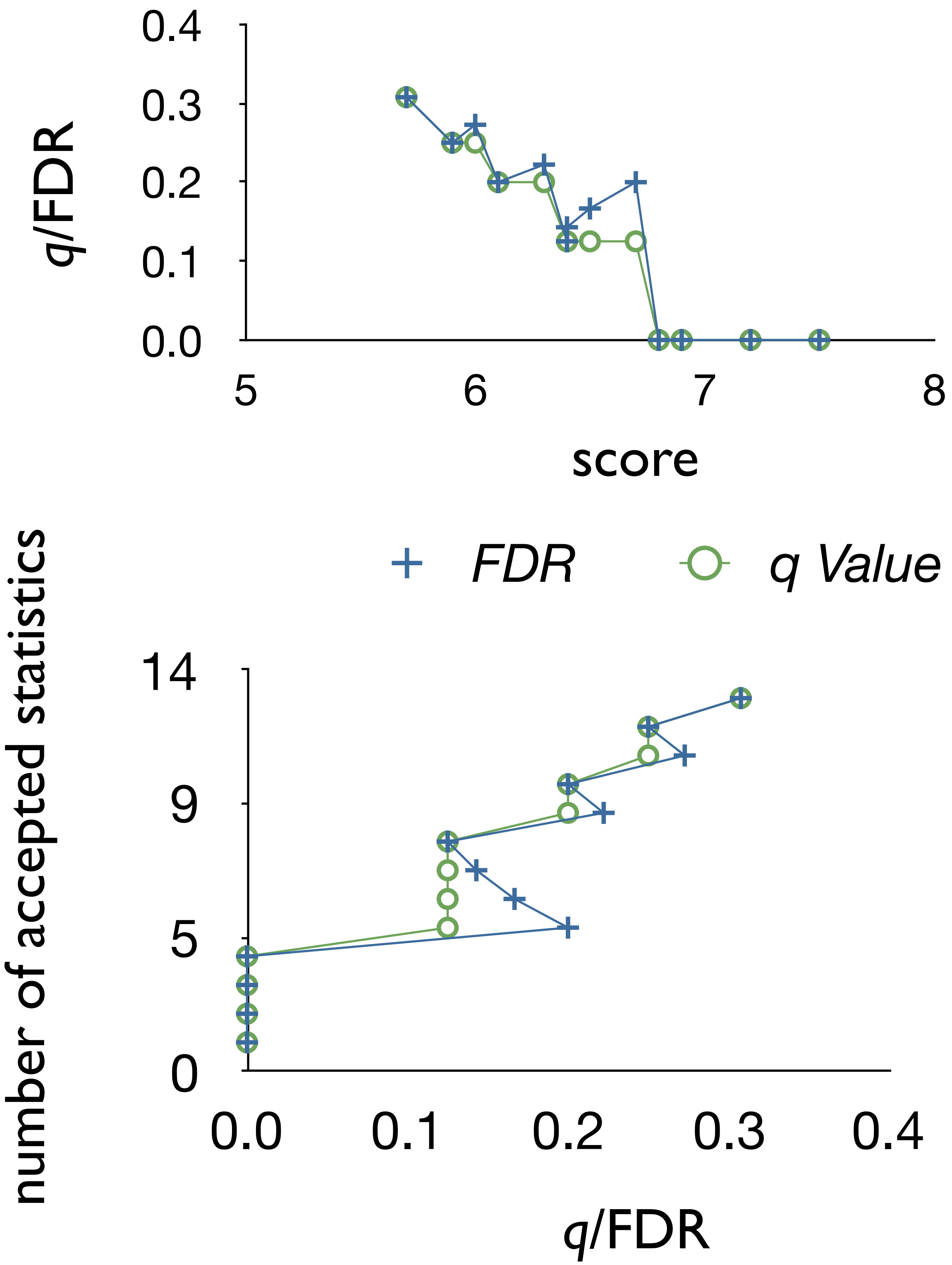




q value

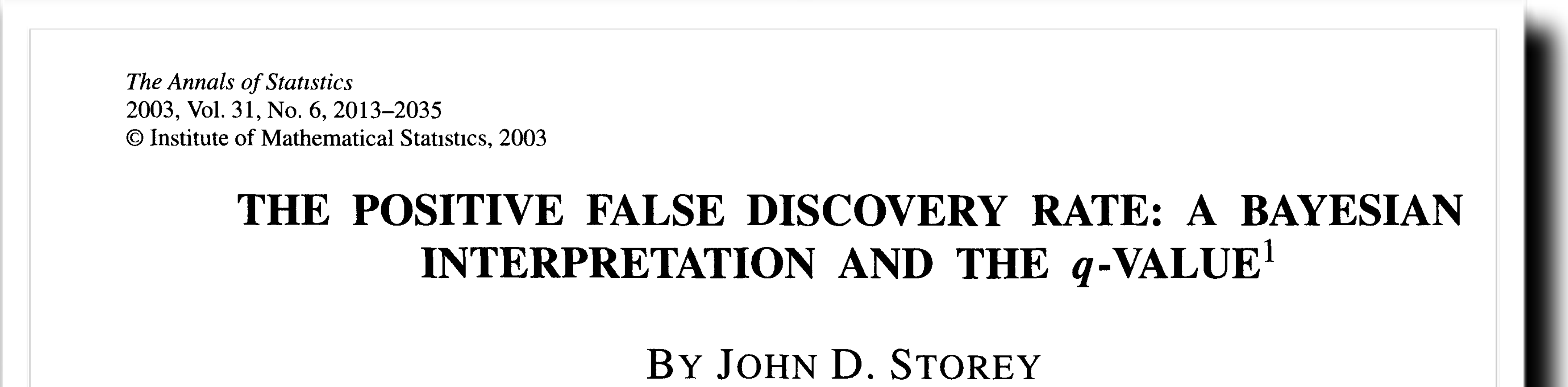
$$q(x)=\min_{x\geq x'}\{\mathrm{FDR}(x')\}$$

score	type
7.5	correct
7.2	correct
6.9	correct
6.8	correct
6.7	incorrect
6.5	correct
6.4	correct
6.4	correct
6.3	incorrect
6.1	correct
6	incorrect
5.9	correct
5.7	incorrect
...	...



Bayesian Interpretation

$$q(t) = \Pr(H = H_0 | p \leq t) = \frac{\Pr(H = H_0)\Pr(p \leq t | H = H_0)}{\Pr(p \leq t)}$$



The Annals of Statıstics
2003, Vol. 31, No. 6, 2013–2035
© Institute of Mathematical Statıstıcs, 2003

THE POSITIVE FALSE DISCOVERY RATE: A BAYESIAN INTERPRETATION AND THE *q*-VALUE[1]

BY JOHN D. STOREY

Bayesian Interpretation

$$q(t) = \Pr(\mathrm{H}=H_0|p\leq t) = \frac{\Pr(\mathrm{H}=H_0)\Pr(p\leq t|\mathrm{H}=H_0)}{\Pr(p\leq t)}$$

$$\widehat{\mathrm{FDR}}(t) = \frac{\hat{\pi}_0 m\cdot t}{S(t)} = \frac{\hat{\pi}_0 m\cdot t}{\#\{p_i \leq t\}}.$$

The Annals of Statıstics
2003, Vol. 31, No. 6, 2013–2035
© Institute of Mathematical Statıstıcs, 2003

THE POSITIVE FALSE DISCOVERY RATE: A BAYESIAN INTERPRETATION AND THE *q*-VALUE[1]

BY JOHN D. STOREY

Bayesian Interpretation

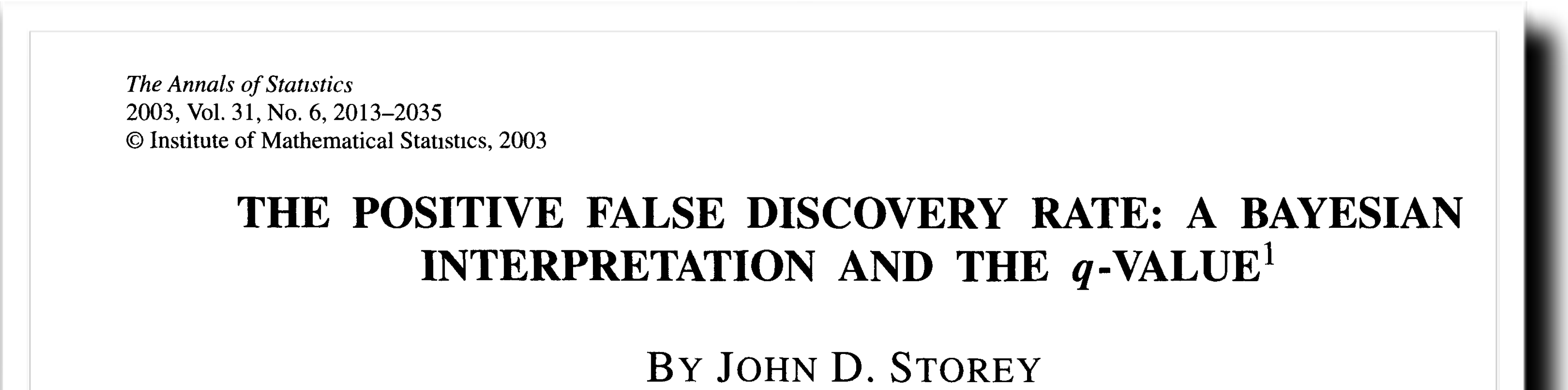
$$q(t) = \Pr(H=H_0|p\leq t) = \frac{\Pr(H=H_0)\Pr(p\leq t|H=H_0)}{\Pr(p\leq t)}$$

$\Pr(H=H_0)$

$\Pr(p\leq t|H=H_0)$

$$\widehat{\mathrm{FDR}}(t) = \frac{\hat{\pi}_0 m\cdot t}{S(t)} = \frac{\boxed{\hat{\pi}_0} m\cdot \boxed{t}}{\#\{p_i \leq t\}}.$$

$\Pr(p\leq t)$



The Annals of Statistics
2003, Vol. 31, No. 6, 2013–2035
© Institute of Mathematical Statistics, 2003

THE POSITIVE FALSE DISCOVERY RATE: A BAYESIAN INTERPRETATION AND THE *q*-VALUE[1]

BY JOHN D. STOREY

FDRs from empirical null models

- If we have an empirical null model, i.e. a mechanism z(y) that models readouts under the null model a *p* value can be estimated as $p(t)=\#\{z(y^i)\geq t\}/(m+1)$

FDRs from empirical null models

- If we have an empirical null model, i.e. a mechanism z(y) that models readouts under the null model a *p* value can be estimated as $p(t)=\#\{z(y^i)\geq t\}/(m+1)$

$$\widehat{\mathrm{FDR}}(t)= \frac{\widehat{\pi_0}\, m\#\{z^i\geq t\}/(m+1)}{\#\{Z^i\geq t\}} \approx \frac{\widehat{\pi_0}\, \#\{z^i\geq t\}}{\#\{Z^i\geq t\}}$$

FDRs from empirical null models

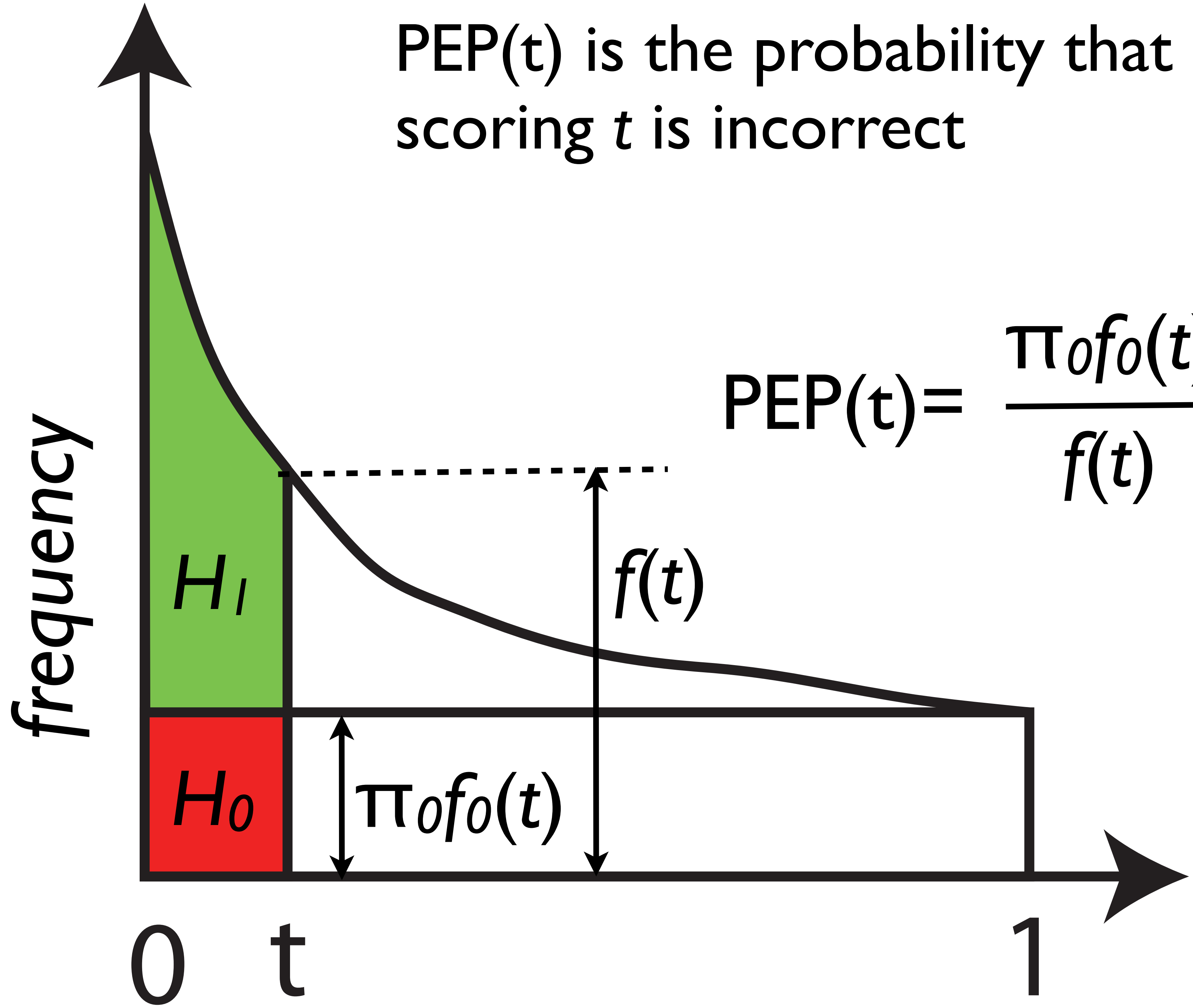
- If we have an empirical null model, i.e. a mechanism z(y) that models readouts under the null model a *p* value can be estimated as $p(t)=\#\{z(y^i)\geq t\}/(m+1)$
- An example: Typically compare difference of trait between sample groups with the ones within a sample group : If $y_H=(y_{H1}, y_{H2})$ and $y_D=(y_{D1}, y_{D2})$ assign significance of $Z=(y_{H1} - y_{D1} + y_{H2}- y_{D2})$ by comparing agains the null model $z=(y_{H1}- y_{H2} + y_{D1}- y_{D2})$

$$\widehat{\mathrm{FDR}}(t)= \frac{\widehat{\pi_0}\, m\#\{z^i\geq t\}/(m+1)}{\#\{Z^i\geq t\}} \approx \frac{\widehat{\pi_0}\, \#\{z^i\geq t\}}{\#\{Z^i\geq t\}}$$

Posterior Error Probability a.k.a. local FDR

PEP(t) is the probability that an identification scoring *t* is incorrect

$$\mathrm{PEP}(t) = \frac{\pi_0 f_0(t)}{f(t)} = \frac{\pi_0 f_0(t)}{\pi_0 f_0(t) + \pi_1 f_1(t)}$$



Control for ...

... FDR or *q* value when you are interested in identifying a sets of significant read-outs

... PEP when you are interested in assessing the quality of a particular read-out

... *p* or *E* value in an experiment rendering one single read-out.